Wave-U-Net for Background Noise Suppression

Signal Processing for AI in Audio

Omri Newman, Adar Cohen, Yuval Sheinin

March 22, 2024

1 Introduction

Noise reduction in speech processing plays a crucial role in improving the quality and intelligibility of audio signals in various applications. It is an essential area that bridges signal processing and machine learning, focusing on enhancing the clarity and comprehension of speech signals compromised by noise, echoes, and various disturbances. Utilizing sophisticated methods, this field aims to better the quality of speech across applications such as telecommunications, digital assistants, sound recordings, hearing devices, and beyond. The pursuit of superior speech quality is a key effort in achieving fluid communication despite the challenges posed by everyday acoustic environments and background noises. The objective of this project is to enhance the clarity of noisy speech by reducing background noise. In this report, we outline our approach to achieving this goal using the Wave-U-Net-PyTorch model and present the results of our experiments.

2 Related Works

In recent years, speech enhancement has garnered significant attention in both research and practical applications. Speech enhancement has traditionally utilized statistical models like hidden Markov models (HMMs) and Gaussian mixture models (GMMs), which leverage the statistical characteristics of speech and noise to differentiate and refine speech signals. An innovative approach within this category is the GFHMM, involving two separate state HMMs and an additional hidden node to model spectral patterns and gain differences [Radfar et al. 2019].

The advent of deep learning has brought transformative advances to speech enhancement. A variety of deep neural networks have been applied to improve speech signals. Techniques such as the Wiener-CNN have been developed to create optimal masks for noise reduction [Mamun el al. 2019]. Another method involves the use of chained generators in a GAN for multi-stage enhancement mapping, progressively improving the quality of noisy input signals [Phan el al. 2020]. Further, a model incorporating a two-stage transformer module (TSTM),

alongside an encoder, a masking module, and a decoder, has been proposed [Wang et al. 2021]. This model excels in extracting both local and global information, applying a mask to the encoded features, and then reconstructing the enhanced speech through decoding.

3 Architecture

The Wave-U-Net-PyTorch model was our model of choice for this project. It is a convolutional neural network architecture designed for end-to-end audio source separation, which addresses the limitations of previously mentioned methods that rely on spectral information and ignore phase details [Stoller et al. 2018]. It operates directly in the time domain, allowing for the modeling of phase information and avoiding fixed spectral transformations. The model uses a multi-scale approach with a structure adapted from the U-Net architecture, processing audio through downsampling and upsampling blocks to capture features at various scales. During downsampling the model uses 1D convolutions to extract features from the input audio with the number of features increasing at each subsequent layer, followed by a decimation step that reduces the time resolution by half. After, the model upsamples the features back to the original resolution, concatenating them with features from the corresponding downsample block, which adds back the fine-grained details lost during downsampling. The final step involves a convolution to output the separated audio sources. Figure 1 includes outlines of the models architecture.

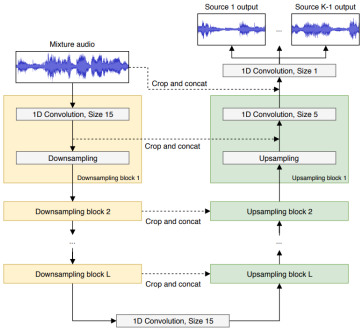

Figure 1. Our proposed Wave-U-Net with K sources and L layers. With our difference output layer, the K-th source prediction is the difference between the mixture and the sum of the other sources.

Block	Operation	Shape
	Input	$(16384, 1)$
DS, repeated for $i = 1, \ldots, L$	$\texttt{Conv1D}(F_c \cdot i, f_d)$	
	$\texttt{Decimate}$	$(4, 288)$
	$\texttt{Conv1D}(F_c \cdot (L+1), f_d)$	$(4, 312)$
US, repeated for $i = L, \ldots, 1$	$\texttt{Upsample}$	
	$\texttt{Concat}$(DS block i)	
	$\texttt{Conv1D}(F_c \cdot i, f_u)$	$(16834, 24)$
	$\texttt{Concat}$(Input)	$(16834, 25)$
	$\texttt{Conv1D}(K, 1)$	$(16834, 2)$

Table 1. Block diagram of the base architecture. Shapes describe the final output after potential repeated application of blocks, for the example of model M1, and denote the number of time steps and feature channels, in that order. DS block i refers to the output before decimation. Note that the US blocks are applied in reverse order, from level L to 1.

Figure 1: Wave-U-Net Architecture

4 Methodology

We conducted several experiments to evaluate the performance of the Wave-U-Net-PyTorch model in reducing noise in speech signals. First we trained the

model on a dataset of clean speech and noisy speech samples by fine-tuning the model, then evaluating its performance using the Signal-to-Noise Ratio (SNR) as a metric. The first experiment checks the robustness of this training procedure on other SNR levels, and the second experiment verifies the performance of the model on specific background noises.

4.1 Creating Dataset

To apply the Wave-U-Net architecture to enhance speech, we utilized the MS-SNSD dataset [Chandan et al. 2019]. This dataset, provided by Microsoft, offers a comprehensive library of clean speech recordings and a diverse array of environmental noise samples, all in .wav format with a 16 kHz sampling rate. Its primary use is in training Deep Neural Network models for the suppression of ambient noise. However, its utility extends to various auditory and speech-related tasks. The dataset includes guidelines for combining clean speech with noise under different SNR ratios and types of noise, allowing for the creation of an extensive noisy speech database tailored to specific application needs. Our model is trained on data generated with a noise level of 20 dB. Some background noises used in modeling include: *NeighborSpeaking, ShuttingDoor, SqueakyChair, Munching, Restaurant, VacuumCleaner, Traffic, WasherDryer, Babble, AirportAnnouncement, Cafe, Metro, Kitchen, CopyMachine, AirConditioner*. The MS-SNSD dataset comes pre-segmented into training and testing subsets. In our study, to ensure an even distribution of data, we implemented a shuffling technique to merge and randomly divide the data into new train and test sets. The train set consists of 2.5 hours of audio samples while the test set consists of roughly 40 minutes of audio samples.

4.2 Training the Model

The Wave-U-Net-PyTorch model expects 5-stem inputs from the MUSDB18 dataset, but we are interested in 2-stem inputs (speech signal, background noise). We first employ a script to process and reformat the audio samples, converting them into acceptable inputs for the Wave-U-Net-PyTorch model, while maintaining the desired 2-stem structure for our use-case.

We then fine-tune the Wave-U-Net-PyTorch model using the reformatted dataset with a starting noise level of 20 dB. The fine-tuning process involves adjusting the model's parameters to optimize its performance for noise reduction in speech. Our baseline assumption for using fine-tuning was that the task of separating vocals and instruments from a song is not far, in terms of the parameters that the model will converge to, from the task of separating speech and background noise.

We experimented with different training strategies, including fine-tuning all layers of the model and modifying specific layers to improve its ability to separate speech from background noise. Although we fine-tuned two models, in the end we report on the results of the model after fine-tuning both downsampling and upsampling blocks.

5 Experiments

The results of our experiments demonstrate the effectiveness of the Wave-U-Net-PyTorch model in reducing noise in speech signals. We observe improvements in SNR performance across different test scenarios, indicating the model's ability to separate speech from background noise accurately. We conduct several experiments to test the model's performance under different conditions, including varying SNRs and testing different background noises. The model consistently demonstrates robust noise reduction capabilities when varying the input SNRs, highlighting its versatility and effectiveness.

5.1 SNR Robustness Check

In this experiment we apply the fine-tuned Wave-U-Net-PyTorch model trained on input signals with noise level of 20 dB on three different SNR groups. We start by generating data with different SNR values (0 dB, 5 dB, 10 dB) and test the model performance on these groups. Each groups test size is roughly five minutes of audio samples. The results presented below indicate a significant increase from each groups baseline SNR, meaning our fine-tuned model succeeds in reducing background noise on groups it was not trained on. In other words, this model is robust to varying levels of background noise.

Table 1: Different SNR values

Input SNR	**Mean SNR (dB)**	**Max. SNR (dB)**	**Min. SNR (dB)**	**Median SNR (dB)**
SNR: 0 dB	14.94	18.43	11.65	14.67
SNR: 5 dB	16.20	18.77	12.46	16.58
SNR: 10 dB	17.04	21.38	13.27	17.02

5.2 Testing Specific Background Noises

Our next experiment includes testing the performance of the fine-tuned Wave-U-Net-PyTorch model on specific types of background noises, also with noise levels of 20 dB. We were able to specify which background noises to include using the MS-SNSD framework to generate data, using Airport Announcements and Typing as the sole source of background noise. Both groups test size are roughly five minutes of audio samples.

We theorized the model would have a harder time differentiating speech signals associated with background noise from the target speech signals. For this reason we expected the model to perform better on non-vocal background noises like Typing. The results below show that while the performance for the Typing group was better than the Airport Announcements as hypothesized, the difference is relatively small. Moreover, the overall performance for these groups were rather poor, we believe this is due to limitations in compute resources and the experiments' test size.

Table 2: Model Performance on Different Background Noises

Noise	Mean SNR (dB)	Max SNR (dB)	Min SNR (dB)	Median SNR (dB)
Airport	18.17	20.26	14.50	18.33
Typing	18.45	20.65	15.08	18.77

6 Results

To evaluate the performance of our fine-tuned model, we measure its SNR performance on the test dataset created by our shuffling technique. We compare the SNR values of the model's output with those of the original recordings to assess the effectiveness of noise reduction. We also compare the SNR results of the pre-trained U-Net model.

Mathematically, SNR is often calculated as the ratio of the power of the signal (the square of the signal's amplitude) to the power of the noise (the square of the noise's amplitude). It is usually expressed in decibels (dB) using the logarithmic scale to provide a more intuitive representation of the difference between signal and noise levels.

Our training and validation MSE loss graph shows that the training loss decreases consistently, indicating the model is learning and improving its performance on the training data. The validation loss also trends downward but appears to stabilize after an initial sharp decline, suggesting the model is generalizing well to unseen data without significant overfitting.

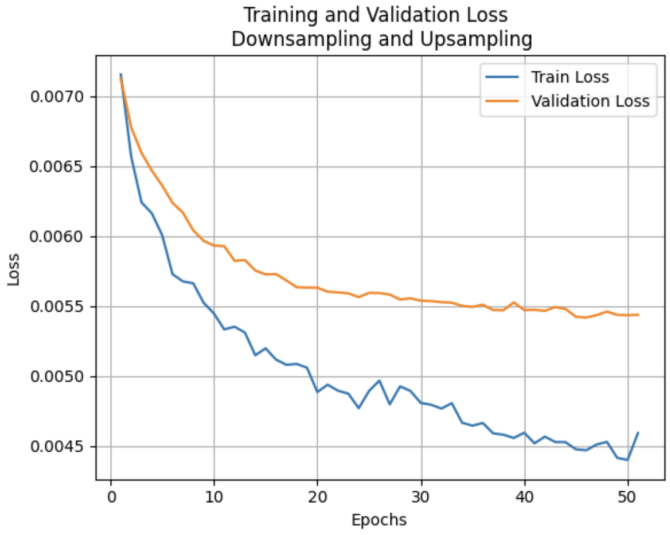

Figure 2: Training and Validation Loss

6.1 Evaluation

We achieved significant improvements in background noise level as seen by the fine-tuned Wave-U-Net-PyTorch model's SNR performance. The model effectively reduced background noise in speech signals from the test set, resulting in higher SNR values compared to the original noisy samples. We expect the results to be even better with more compute resources and longer run-time.

SNR	Mean SNR (dB)	Max SNR (dB)	Min SNR (dB)	Median SNR (dB)
SNR 20	23.43	26.87	20.08	23.66

6.2 Comparison with Pre-Trained Wave-U-Net Model

We compare the results obtained from the pre-trained Wave-U-Net model using the SNR metric on the test set. These results demonstrate the effectiveness of our model in terms of noise reduction and signal quality enhancement, supporting our decision to fine-tune it.

SNR	Mean SNR (dB)	Max SNR (dB)	Min SNR (dB)	Median SNR (dB)
SNR 20	18.41	23.73	9.23	19.23

7 Conclusion

Our project successfully demonstrates the effectiveness of using the Wave-U-Net-PyTorch model for noise reduction in speech processing. The results highlight the potential of deep learning based approaches for addressing real-world signal processing challenges. While we achieve significant improvements in noise reduction, we identify some limitations and aspects that can be further explored:

Robustness to Various Noise Types: As seen in our experiment, the model's performance may vary depending on the type of background noise present in the audio signal. Future research could focus on enhancing the model's robustness to different types of noise.

Generalization to Other Datasets: Our experiments were conducted solely on the MS-SNSD dataset. It would be beneficial to evaluate the performance of the model on other datasets to assess its generalization capabilities.

Real-time Processing and Deployment: Investigating techniques to optimize the model for real-time noise reduction is essential for practical deployment in real-world scenarios such as live communication systems or audio recording devices. This includes addressing challenges with computational resources and latency constraints.

8 Supplementary Material

The link to our code, final presentation, recordings, and all other supplementary material can be found here.

References

[1] Stoller, D., Ewert, S. and Dixon, S., 2018. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185.

[2] Radfar, M. H., Dansereau, R. M., Wong, W. (2019). Speech Separation Using Gain-Adapted Factorial Hidden Markov Models. ArXiv. /abs/1901.07604

[3] Mamun, N., Khorram, S., Hansen, J. H. (2019). Convolutional Neural Network-based Speech Enhancement for Cochlear Implant Recipients. ArXiv. /abs/1907.02526

[4] Phan, H., McLoughlin, I. V., Pham, L., Chén, O. Y., Koch, P., De Vos, M., Mertins, A. (2020). Improving GANs for Speech Enhancement. ArXiv. https://doi.org/10.1109/LSP.2020.3025020

[5] Wang, K., He, B., Zhu, W. (2021). TSTNN: Two-stage Transformer based Neural Network for Speech Enhancement in the Time Domain. ArXiv. /abs/2103.09963

[6] Chandan KA Reddy et al., "A scalable noisy speech dataset and online subjective test framework," arXiv preprint arXiv:1909.08050, 2019.

[7] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 323–332, 2017.